A Closer Look A Seattle Census Tract Data: Primary Components and Cluster Analysis

Niva Razin

December 2021

1 Introduction

1.1 Data

The data set used in this paper was published by the City of Seattle GIS Program; I accessed it through the U.S. Government's open data website, *data.gov* [2]. The data set was used by the City of Seattle Office of Planning & Community Development to develop the Racial and Social Equity Composite Index "to aid in the identification of City planning, program and investment priorities" based on "where priority populations make up relatively large proportions of neighborhood residents" [3].

Each observation ($n = 134$) in this data set corresponds to a unique census tract in Seattle, Washington. Census tracts are small, relatively permanent geographical regions that are statistical subdivisions of a county. Each census tract has a population between 1,200 and 8,000, with an average of about 4,000 [1].

This data set includes over two dozen variables related to the racial/ethnic composition (e.g., percent people of color), nation-of-origin (e.g., percent foreign born), health (e.g., prevalence of obesity, asthma, and diabetes), level of education (e.g., percent with less than bachelor's degree), and socioeconomic standing (e.g., socioeconomic percentile) of the inhabitants of each census tract. Subsets of these variables are used for ease of analysis, as explained below (**2.1**).

Finally, one observation was omitted from the analyses due to missing data; 134 unique census tracts were included in the analysis.

1.2 Motivation & Goals

The Racial and Social Equity Composite Index developed by the City of Seattle from this data set assigns a single Composite Index to each observed census tract. This index is then used to bin the tracts into quintiles of priority/disadvantage (i.e., "Lowest", "Second Lowest", "Middle", "Second Highest", and "Highest"). Single summary statistics like this index are convenient

and easy to interpret. However, as with most summary statistics, it may obscure more nuanced relationships among the data. Moreover, the Composite Index may be used to determine the allocation of millions of dollars of funds to different Seattle census tracts over the next decade. Thus, it is critical that the level of disadvantage/priority and the grouping of census tracts be deeply understood and that there be multiple means of characterizing these levels/groupings beyond a single summary statistic.

This paper aims to more closely analyze the data to elucidate potential relationships among variables and clusters of similar tracts. In particular, we seek to investigate the following questions:

1. How can we succinctly summarize the priority / disadvantage level a tract? Can we better, more clearly capture the socioeconomic, racial/ethnic makeup, etc. of these tracts outside of the Composite Index?

2. Which census tracts are most similar? Is there another way to conceptualize groupings of tracts by similar socioeconomic standing, racial/ethnic makeup, etc. outside of the Composite Index?

To investigate (1), we will perform principal components analysis on a subset of the variables in the original data set. We will use this same subset to perform cluster analysis on the data for (2).

2 Results & Analysis

2.1 Variables

We will use the following variables associated with each census tract for the analyses:

- POVERTY: Percentile based on share of population with income below 200% of the poverty level
- DEGREE: Percentile based on share of population with level of education less than a bachelor's degree
- POC: Percentile based on share of population that is people of color
- FOREIGN: Percentile based on share of population that was not born in the U.S.
- MENTAL: Percentile based on share of adults with mental health "not good"
- OBESE: Percentile based on share of adults that is obese

These variables were selected to attempt to capture the many dimensions of demographic information represented in the data: there is one variable associated with socioeconomic standing (POVERTY), one with educational attainment (DEGREE), one with racial/ethnic make-up (POC), one with nation-of-origin (FOREIGN), and two with health, one mental and one physical (MENTAL and OBESE, respectively).

Note that the percent associated with each of the above variables (e.g., percent of population that is people of color) are also provided in the data set. Here, we use the percentiles instead of the percentages because the percentile values are more evenly distributed between 0 and 1, which is convenient for our analyses.

2.2 Principal Components Analysis

Most of the variables are highly correlated, with a few pairs showing little to moderate correlation (**4.1**). Among the most highly correlated variables are MENTAL and POVERTY ($r = 0.880$), FOREIGN and POC ($r = 0.876$), and DEGREE and MENTAL($r = 0.813$). Among the least highly correlated variables are OBESE and FOREIGN ($r = 0.264$), OBESE and POVERTY ($r = 0.437$), and OBESE and POC ($r = 0.474$).

We will use the covariance matrix for PCA because all these variables have the same variance and range as they are percentile variables (**4.1**).

We perform PCA with all variables and find that the first principal component, PC1, accounts for 0.7309 of variance in the data. PC1 and PC2 account for a cumulative 0.8815 of variance; PC2 only contributes an additional 0.1505, which is less than 1/6. So, PC1 should suffice (**4.2**).

PC1 appears to be an average of the six variables, with a bit less weight given to OBESE. It is also highly correlated with all variables except OBESE, with which it has a moderate correlation. PC2 largely represents a weighted difference in the OBESE and FOREIGN variables (**4.2**).

Based on the relatively low correlation coefficients between OBESE and the other variables (**4.1**), we perform PCA again, omitting this variable (**4.3**). We find that the proportion of variance explained by PC1 then jumps to 0.8020. Again, PC2 contributes less than 1/5 in explanation of variance, so we omit it; PC1 will suffice. PC1 also again appears to be an average of the five variables, though this time more evenly weighted across the included variables. PC1 scores are highly correlated with all five variables, too. PC2 largely represents a weighted difference in the FOREIGN and DEGREE (and perhaps MENTAL) variables.

Taken together, these PCA analyses suggest a significant amount of the variance in the data can be captured by a single linear combination of five percentile variables, PC1 of PCA on all but OBESE variables. Thus, $PC_1 = 0.45 * POVERTY + 0.420 * DEGREE + 0.469 * POC + 0.424 * FOREIGN + 0.466 * MENTAL$ offers another metric by which we can capture the data *and* see how it is being calculated. Thus, we might use the PC1 scores alongside

the Composite Index to add another dimension/metric to our rank of level of priority/disadvantage of tracts.

Our ability to omit OBESE suggests it may not be the most informative physical health indicator variable. That is, obesity prevalence may vary little with other demographic variables across Seattle's census tracts. Instead of omitting this physical health indicator, we may replace it with another physical health indicator variable in future analyses, such as the percentile based on share of adults that have asthma, which is also included in this data set.

2.3 Cluster Analysis

Next, we use non-hierarchical k-means clustering to aggregate the census tracts into groups. We make five clusters and generate a pairplot with the clusters as colorings (**4.4**). We can see positive linear trends among most of these variables. Additionally, the colorings appear to distinguish segments of these plots fairly well, though there is some overlap among them.

We used five clusters because these are how many bins the City of Seattle used for their Composite Index. The k-means clustering appears to roughly group tracts in a similar manner as the Composite Index quitiles, i.e., with tracts in the highest percentiles grouped together and likewise for those in the lowest percentiles.

Finally, we see a relatively high level of compactness of the groups (between_SS / total_SS = 75.7%), indicating a good fit of clusters for $k = 5$.

We chose non-hierarchical clustering here because the evolution of the formation of groups is not very important to us. Additionally, a dendrogram is not meaningfully interpretable with a sample size $n = 134$ and arbitrary labels (i.e., census tract codes).

Overall, k-means clustering appears to provide a good fit to the data and an alternative grouping system to the bins formed using the Composite Index.

3 Conclusions

Multivariate methods add a rich dimension of analysis to the Seattle census tract data set and Composite Score. PCA provided us with principal components, or variables with easily interpretable linear combination structures of the original variables that capture nearly as much variance as these original variables. Future work could compare the PC1 scores discussed in **2.2** to the Composite Index values to see how these metric are similar and different.

Additionally, cluster analysis provided us with an alternate way to group Seattle census tracts. In turn, these clusters might be used to find underlying similarities and issues associated with regions of higher or lower disadvantage. These clusters could be compared with the quintile groups determined by the City's Composite Index, offering another means to understand and aggregate similar tracts.

Finally, future analyses could explore PCA and cluster analysis using different subsets of this data set; this paper only used a small subset of the variables available.

4 Supporting Material

4.1 EDA Code

```
all_data<-read.table('sub_data.csv', header = TRUE, sep=',')

# look at numerical data (not census tract IDs and indices)
data <-all_data[,4:9]
cor(data)

##           POVERTY    DEGREE       POC   FOREIGN    MENTAL     OBESE
## POVERTY 1.0000000 0.6631275 0.8035896 0.7101264 0.8804195 0.4365189
## DEGREE  0.6631275 1.0000000 0.7310717 0.5819385 0.8127267 0.7260965
## POC     0.8035896 0.7310717 1.0000000 0.8758113 0.7977803 0.4744492
## FOREIGN 0.7101264 0.5819385 0.8758113 1.0000000 0.6463905 0.2639741
## MENTAL  0.8804195 0.8127267 0.7977803 0.6463905 1.0000000 0.6222263
## OBESE   0.4365189 0.7260965 0.4744492 0.2639741 0.6222263 1.0000000

# these are percentile variables, hence their identical variances
apply(data, 2, sd)

##   POVERTY    DEGREE       POC   FOREIGN    MENTAL     OBESE
## 0.2919017 0.2919017 0.2919017 0.2919017 0.2945721 0.2944069
```

4.2 PCA Code: All Variables

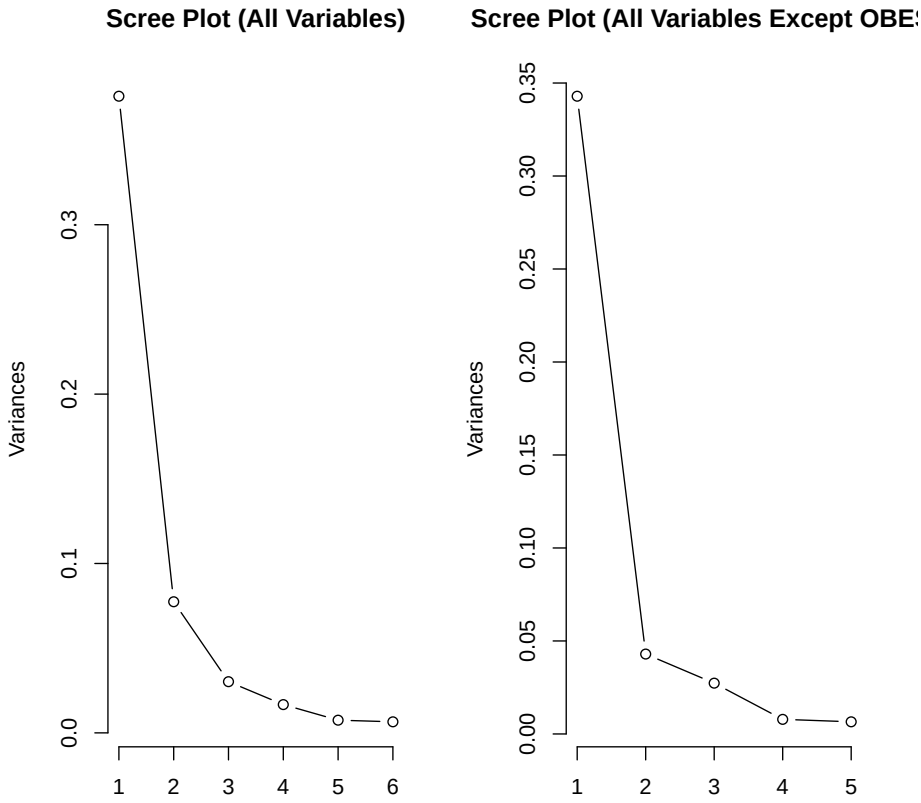
```
# PCA with all variables
data.pca<-prcomp(data, scale = F)
data.pca

## Standard deviations (1, .., p=6):
## [1] 0.61310165 0.27824605 0.17386634 0.12922557 0.08657401 0.08084719
##
## Rotation (n x k) = (6 x 6):
##                PC1         PC2        PC3        PC4        PC5         PC6
## POVERTY 0.4232743 -0.20371776 -0.6179770  0.2053765 -0.2834878 -0.52430670
## DEGREE  0.4188781  0.30756922  0.2015894 -0.7796709 -0.1327763 -0.25256241
## POC     0.4391094 -0.26156265  0.2769519  0.1473492  0.7672247 -0.22742037
## FOREIGN 0.3833764 -0.50728853  0.4903364  0.1359722 -0.5178411  0.26192216
```

```
## MENTAL   0.4499109  0.07164897 -0.4567571 -0.1156434  0.1720124  0.73543042
## OBESE    0.3211941  0.73007534  0.2273915  0.5443897 -0.1249948 -0.01155712

data.pca.scores<-predict(data.pca)

summary(data.pca)

## Importance of components:
##                            PC1    PC2     PC3     PC4     PC5     PC6
## Standard deviation     0.6131 0.2782 0.17387 0.12923 0.08657 0.08085
## Proportion of Variance 0.7309 0.1505 0.05878 0.03247 0.01457 0.01271
## Cumulative Proportion  0.7309 0.8815 0.94024 0.97272 0.98729 1.00000

# correlation between percentile variables and PCA scores
cor(data, data.pca.scores)

##               PC1         PC2        PC3         PC4         PC5          PC6
## POVERTY 0.8890328 -0.19418751 -0.3680876  0.09092064 -0.08407856 -0.145215748
## DEGREE  0.8797992  0.29318062  0.1200733 -0.34516215 -0.03937960 -0.069951498
## POC     0.9222923 -0.24932632  0.1649618  0.06523184  0.22754824 -0.062987980
## FOREIGN 0.8052324 -0.48355674  0.2920607  0.06019523 -0.15358451  0.072543841
## MENTAL  0.9364128  0.06767797 -0.2695933 -0.05073149  0.05055402  0.201843544
## OBESE   0.6688858  0.68999931  0.1342894  0.23895183 -0.03675628 -0.003173704
```

4.3 PCA Code: All Variables Except OBESE

```
# PCA again but without OBESE
data_but_obese.pca <- prcomp(~POVERTY+DEGREE+POC+FOREIGN+MENTAL, data=data, scale = F)
data_but_obese.pca

## Standard deviations (1, .., p=5):
## [1] 0.58559087 0.20729618 0.16528492 0.08861698 0.08086208
##
## Rotation (n x k) = (5 x 5):
##               PC1         PC2        PC3        PC4        PC5
## POVERTY 0.4534428 -0.04881833  0.6695591 -0.2788971  0.5156679
## DEGREE  0.4208432 -0.54418077 -0.6288296 -0.2561087  0.2563982
## POC     0.4690964  0.30134833 -0.1623495  0.7759040  0.2464821
## FOREIGN 0.4238691  0.68093324 -0.2234580 -0.4793679 -0.2773768
## MENTAL  0.4664378 -0.38340992  0.2827952  0.1574935 -0.7284609

data_but_obese.pca.scores<-predict(data_but_obese.pca)

summary(data_but_obese.pca)
```

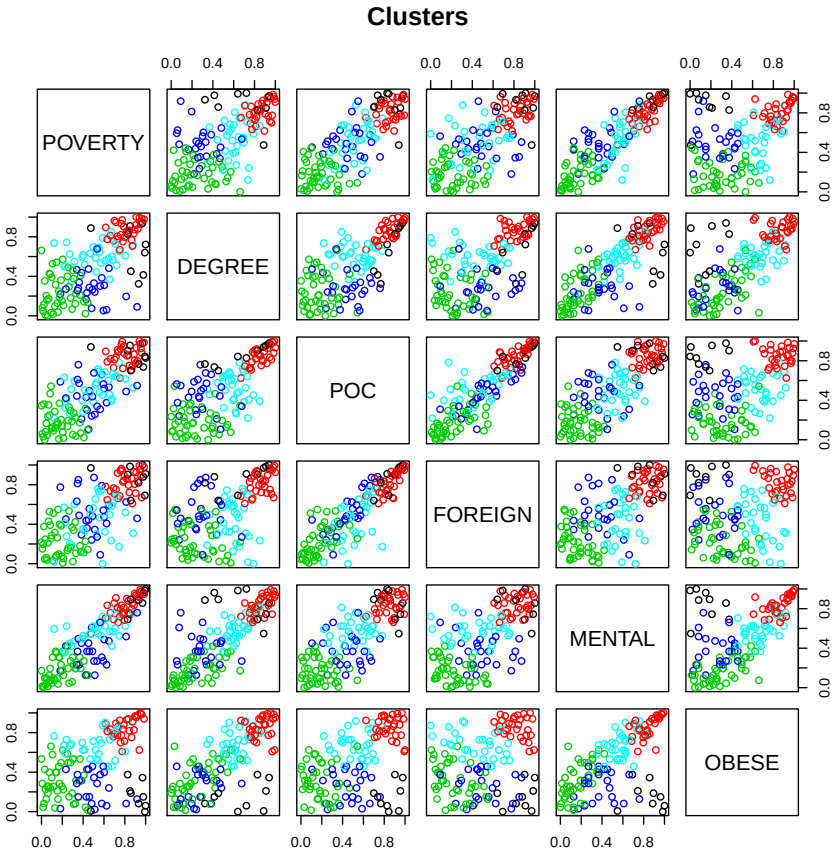
```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5
## Standard deviation     0.5856 0.2073 0.16528 0.08862 0.08086
## Proportion of Variance 0.8020 0.1005 0.06389 0.01837 0.01529
## Cumulative Proportion  0.8020 0.9024 0.96634 0.98471 1.00000

# correlation between percentile variables and PCA scores
cor(data[,1:5], data_but_obese.pca.scores)

##               PC1        PC2         PC3         PC4         PC5
## POVERTY 0.9096623 -0.0346687  0.37912775 -0.08466897  0.14284939
## DEGREE  0.8442635 -0.3864541 -0.35606524 -0.07775077  0.07102695
## POC     0.9410653  0.2140048 -0.09192795  0.23555283  0.06828001
## FOREIGN 0.8503338  0.4835699 -0.12652970 -0.14552890 -0.07683842
## MENTAL  0.9272490 -0.2698131  0.15867686  0.04737923 -0.19996753

# Scree Plots
par(mfrow=c(1,2))

plot(data.pca, type="line", main ="Scree Plot (All Variables)")

plot(data_but_obese.pca, type="line",
    main ="Scree Plot (All Variables Except OBESE)")
```



4.4 Cluster Analysis Code

```
cl <- kmeans(data, 5, nstart = 100)
cl

## K-means clustering with 5 clusters of sizes 9, 30, 41, 24, 30
##
## Cluster means:
##     POVERTY    DEGREE       POC   FOREIGN    MENTAL     OBESE
## 1 0.8760000 0.6797778 0.8501111 0.8666667 0.8407778 0.1573333
## 2 0.8152000 0.8656000 0.8523333 0.8207667 0.8385667 0.8541333
## 3 0.1802439 0.2414390 0.1803902 0.2271707 0.1664634 0.3075122
## 4 0.4932917 0.2872500 0.4537917 0.5124583 0.3767500 0.2528333
## 5 0.5165667 0.6062333 0.5186000 0.4343333 0.5728000 0.6754667
```

```
##
## Clustering vector:
##   [1] 4 4 5 5 3 4 1 4 4 5 3 4 3 2 2 3 2 3 2 5 2 3 2 5 5 4 3 2 5 2 5 2 5 2 3 3 3
##  [38] 3 3 5 2 3 4 2 1 2 1 5 5 4 3 4 3 5 3 3 3 1 4 3 5 2 2 2 2 5 2 2 1 2 5 5 2 5
##  [75] 4 2 3 3 5 3 4 5 2 1 2 3 3 5 5 4 1 3 4 1 2 3 5 5 3 3 3 3 4 5 4 4 4 3 3 2 2
## [112] 2 3 5 3 3 3 5 3 3 3 5 5 3 2 4 4 4 2 1 5 3 4 4
##
## Within cluster sum of squares by cluster:
## [1] 1.197994 1.780425 5.489564 3.849125 4.324043
##  (between_SS / total_SS =  75.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"

plot(data, col = cl$cluster, main="Clusters")
```



References

[1] *Glossary*. Oct. 2021. URL: https://www.census.gov/programs-surveys/geography/about/glossary.html#par_textimage_13.

[2] *Racial and social equity composite index*. Sept. 2021. URL: https://catalog.data.gov/dataset/racial-and-social-equity-composite-index.

[3] City of Seattle GIS Program. *Racial and Social Equity Composite Index*. URL: https://data-seattlecitygis.opendata.arcgis.com/datasets/SeattleCityGIS::racial-and-social-equity-composite-index/about.